Theory and Decision 69.4 (2010): 569-586

DOI: http://dx.doi.org/10.1007/s11238-009-9132-8

Guilt and Shame: An Axiomatic Analysis

Raúl López-Pérez

October 2008

Abstract

Using the machinery of Game Theory, this paper analyzes how shame and guilt affect preferences. Based on abundant psychological literature, we posit that the preference ordering of someone who can feel shame (or guilt) must satisfy a number of axioms and prove that it can be represented by a particular utility function. Understanding how shame and guilt work is important to explain why people respect social norms and exhibit prosocial behavior, many times contrary to their material interest.

Keywords: Guilt, Inferiority Feelings, Norms, Reciprocity, Self, Shame. JEL classification numbers: C70, C72, D63, D64, Z13.

1 Introduction

Many social researchers -Arrow (1974), Elster (1989)- have argued that *social norms* are crucial for the attainment of social order. Partly for this but due also to an increased interest on other-regarding preferences, economists are paying a growing attention to issues like norm enforcement and the effect of norms on preferences and behavior - Becker (1996), Conlin et al. (2003), Fehr and Fischbacher (2004), Akerlof and Kranton (2005). In addition, emotions are also attracting much attention, as the analysis of patients with specific brain lesions suggests that they play a key role in explaining pro-social behavior. In particular, Damasio (1994) argues that lack of emotions results in asocial behavior -sociopaths seem to be the extreme case in this regard.[1]

Guilt and *shame* (two examples of what social psychologists call self-conscious emotions; see Tangney and Dearing, 2002) seem particularly important in explaining why people comply with norms. Indeed social researchers like Emile Durkheim and Talcott Parsons long ago contended that people often comply with internalized norms in order to avoid feeling ashamed or guilty, and this line of reasoning is consistent with the experimental evidence from Bosman and van Winden (2002) or the evidence from Beer et al. (2003), which report that lesions in the orbitofrontal brain region are highly correlated with *both* abnormal functioning of self-conscious emotions and inappropriate behavior - e.g., orbitofrontal patients tend to include sexually intimate (and unasked) details when describing past emotional experiences to strangers.

Our paper uses the standard apparatus of rational preferences and games to analyze how guilt and shame affect preferences when someone has internalized a norm. To our knowledge, our paper is the first to formally study this issue, something that seems important in order to organize the available evidence and promote further theoretical research. In this respect, and although our paper is related to recent utility models on guilt and shame -Battigalli and Dufwenberg (2007) and Tadelis (2008)-, it distinguishes from them in that they do not incorporate norms into the analysis and do not provide an axiomatic study of the preferences.

The influence of shame and guilt on norm compliance seems important to understand numerous social phenomena which are difficult to explain if one assumes that *everybody* is selfish, like contributing to charities, cooperating with NGOs, helping strangers in distress, abstaining from being adulterous, voting in general elections, maintaining a vegetarian diet out of respect for animal life, paying taxes, keeping the streets clean, saving water

[1] Mealy (1995, p. 523) notes that "sociopaths, who comprise only 3-4 percent of the male population and less than 1 percent of the female population, are thought to account for 20 percent of the United States prison population and between 33 percent and 80 percent of chronic criminal offenders."

during droughts, avoiding free-riding at public transport, visiting cemeteries to honor deceased family or friends, participating in demonstrations, and recycling glass and paper. Or why should firms spend huge amounts of money in efforts to improve society and safeguard the environment, guided by ideas of Corporate Social Responsibility (CSR),[2] if *nobody* cared about that?

The rest of the paper proceeds as follows. Section 2 introduces norms into a game-theoretical analysis and illustrates the concept with several examples. In section 3 we model the preferences of someone who has internalized a norm, and prove that any rational preference ordering satisfying four sensible axioms can be represented by a precise utility function. Importantly, we have shown in López-Pérez (2008) that a model based on this utility function (and other hypotheses) can explain a large number of robust experimental facts (including some that other models of other-regarding preferences have problems to explain). Hence, our formal study here is also consistent with that experimental evidence and, in particular, with the mounting evidence showing that people tend to cooperate conditionally or reciprocally -see Fehr and Gächter (2000) for a survey. In addition, our model here is also in line with the psychological literature on guilt and shame, which we review in section 4 in some detail. We conclude in section 5 by discussing future lines of research.

2 Introducing Norms in Games

This section introduces norms, which are the building block of our analysis, into a game-theoretical framework. Given any game in extensive form, let h denote an information set and $A(h)$ denote the set of actions available at h.

Definition 1 *A norm is a nonempty correspondence* $\Psi : h \rightarrow A(h)$ *applying on any* h.

Intuitively, a norm is an "ought" statement, that is, a rule indicating how one should behave. More precisely, one can think of the selected actions at h as the choices commended by the norm, while non-selected actions constitute deviations or transgressions. Note also that defining a norm as a nonempty correspondence means that it always commends at least one choice at any conceivable situation. This might seem too demanding as most actual norms refer to very specific situations -the norm to wear black in funerals, for instance, is silent about how one should behave in a wedding party, or when travelling in the subway. Nevertheless, one can accommodate this kind of norms within our defin-

[2] For a survey on CSR, see *The Economist*, January 19th 2008.

ition by simply imposing $\Psi(h) = A(h)$ when convenient -e.g., the norm to wear black in funerals applies to weddings as well, although it does not restrict choice there at all.

We have chosen definition 1 because it is simple, general (we see no fundamental reason to restrict the range of application of the model), and because we want to distinguish clearly between norms and the factors that might explain compliance with them. Furthermore, it is in line with some classical definitions of norms, like those of Parsons (1937) and Homans (1961).[3] However, we stress that definition 1 differs from some others that appear in the literature. This is unavoidable because, as Horne (2001, 3) notes "[...] scholars disagree about what norms are. To complicate matters, they use a variety of terms -custom, convention, role, identity, institution, culture, and so forth- to refer to concepts that are similar to or overlap with notions about norms. Furthermore, the word has various meanings depending on the focus of the researcher." For instance, definition 1 diverges from an extended usage in Game Theory, which defines a norm as a behavioral regularity or equilibrium prediction of actual play -as in Kandori (1992) or Voss (2001). Contrary to that, our model allows that players deviate in equilibrium from a norm: The enforcement of a norm depends on the players' preferences (to be analyzed in the following section) and the specific game being played. Our definition also contrasts with Posner's (1997, 365): "a rule that is neither promulgated by an official source, such as a court or a legislature, nor enforced by the threat of legal sanctions, yet is regularly complied with". According to our definition (which is indeed very general), a law is also a norm.

To clarify matters, we also note that much literature refers to social norms, and not (simply) norms. In this respect, we could define a *social* norm as a norm that numerous people have internalized (see next section), a definition that is somehow similar to others present in the literature. Thus, Elster (1989, 99) define social norms as prescriptions that are "(a) shared by other people and (b) partly sustained by their approval and disapproval", and Fehr and Fishbacher (2004, 185) as "standards of behavior that are based in widely shared beliefs about how individual group members ought to behave in a given situation". Finally, Becker (1996, 225) apparently refer to social norms when stating that "Norms are those common values of a group which influence an individual's behavior through being internalized as preferences."

Since definition 1 is very flexible, it allows us to represent a large range of normative and ethical intuitions. To illustrate this point, we will dedicate the rest of this section to provide numerous formal examples of norms. For this, let $N = \{1, .., n\}$ denote the set of players of the game, z denote a terminal node, and $x_i(z)$ player i's material payoff at z. This material payoff can be thought as player i's monetary payment or, more generally,

[3] See Horne (2001) for a review of the sociological literature on norms.

as a *cardinal* measure of the utility that i gets from consumption and leisure through the history of z. Crucially, material payoffs and utility payoffs need not coincide, as utility may be influenced by norms and emotions -more on this later.

In addition, let $X(h)$ denote the set of all material allocations $x = (x_1, x_2, .., x_n)$ that succeed information set h, and $F : \mathbb{R}^n \to \mathbb{R}$ denote any continuous function. We say that vector $x \in X(h)$ is an F-fairmax allocation of h if it maximizes function F over the set $X(h)$, whereas an action $a \in A(h)$ is an F-fairmax action if it points towards at least one F-fairmax allocation of h.[4]

Given any function F, consider a correspondence that selects all the F-fairmax actions of h. If we interpret function F as a social welfare function, we can view this correspondence as a norm of distributive justice, as it commends to move towards a fair outcome -i.e., an F-fairmax allocation. Of course, there exist infinite such norms, as there exist infinite functions F. A prominent example, though, might be the *efficiency norm*, which corresponds to the following function:

$$F(x) = \sum_{i \in N} x_i. \tag{1}$$

As another example one may cite the Maximin or *Rawlsian norm*, which corresponds to function

$$F(x) = Min\{x_1, ..., x_n\}. \tag{2}$$

To illustrate how these two norms select actions, consider an agent B who has to choose one of the following three (B, other) allocations of money payoffs: (7, 0), (0, 10) and (3, 3). In this simple decision problem, the efficiency norm commends B to choose (0, 10) whereas the Rawlsian norm prescribes to choose (3, 3).

These simple norms share two features. First, they are *history-independent*: They implicitly state that all agents, *including previous transgressors*, should be equally considered at any moment. One could relax this feature by assuming that function F depends on previous history and thus construct *reciprocal norms*, that is, norms whose prescriptions are conditional on whether others respected the norm *previously*. As an illustration,

[4]The set $X(h)$ might not be compact -i.e., closed and bounded- and, therefore, function F might not have a maximum on $X(h)$. But this does not seem to pose a big problem. First, having an unbounded set $X(h)$ appears very unlikely in real world problems. Second, in case $X(h)$ is bounded but not closed, one may define an $F-$fairmax allocation of h by means of the following procedure. First, find the smallest closed set containing $X(h)$. Second, find the element(s) of that set that maximize F. Third, choose a positive number ε and consider a closed ε-disk with center any of the previous elements. Then, any vector(s) in $X(h)$ within such ε-disk is defined as an $F-$allocation of h. In what follows, however, we will ignore these subtleties and assume that $X(h)$ is always compact.

consider a *reciprocal-efficiency norm* commending to choose an $F[R(t)]$-fairmax action at any decision node t of a perfect information game,[5] where

$$F[R(t)] = \sum_{i \in R(t)} x_i \tag{3}$$

and $R(t)$ is the set of players that respected the norm in the history of t. To apply this norm, we posit that $R(t_0) = N$ at the initial decision node t_0. Consequently, this norm selects at t_0 the same actions as the efficiency norm (1). If t_1 is a node immediately succeeding t_0 and the action leading from t_0 to t_1 is selected by the norm, it follows that $R(t_1) = N$ and hence the norm selects again the same choices at t_1 as norm (1). In contrast, if node t^* succeeds t_0 but the action leading from t_0 to t^* is not selected by the norm, $R(t^*)$ then contains all players except the first mover (since he deviated when moving from t_0 to t^*). One may proceed in an analogous way to describe set $R(\cdot)$ for the remaining decision nodes, and then apply function (3) to determine what actions are selected by the norm at each node. Intuitively, this norm commends to maximize at any moment the sum of material payoffs of all players who previously respected the norm. Alternative reciprocal norms -like a reciprocal Rawlsian norm- may be obtained in a similar fashion.

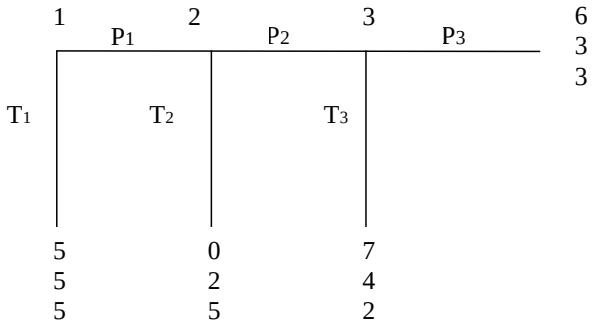

Figure 1: A Three-Players Game.

To show the differences between the efficiency norm of function (1) and its reciprocal version -function (3)- consider the three-player game tree depicted at Figure 1, which displays only money payoffs. In this sequential game, each player may pass (P) the turn to the following player (if any) or terminate the game (T). Both norms suggest player 1 to choose T_1 in order to maximize the social surplus. If player 1 deviates and chooses P_1, in contrast, the efficiency norm commends to move towards allocation (7, 4, 2) and hence choose P_2, while its reciprocal version suggests to choose T_2 -notice that $R(t)$ is then equal

[5] We later suggest how to extend these ideas to imperfect information games.

to $\{2,3\}$. Finally, the efficiency norm commends to play T_3 if player 2 moved P_2 while the reciprocal norm suggests in this case to move P_3. In other words, the efficiency norm and its reciprocal version respond in different ways to a previous violation of what they prescribe. The efficiency norm asserts that all agents, including previous transgressors, are on the same footing and should enter equally into one's ethical calculations at any moment. The reciprocal version, on the opposite, only considers non-transgressors as deserving good treatment.

The idea of reciprocity is closely linked to that of *retributive justice*, that indicates how transgressors *should* be punished. While the reciprocal efficiency norm of function (3) recommends not to be kind towards transgressors, one might think of more severe norms commending to punish transgressors, maybe within some limits. Consider, for instance, the following case

$$F = \sum_{i \in R(t)} x_i - (\max_{j \notin R(t)} x_j - \min_{i \in R(t)} x_i).$$

This norm commends to punish the *best-off* deviator although the amount of the sanction is bounded: The deviator should never end up poorer than the *worst-off* norm follower. One may think of related examples, like *Lex Talionis* ('an eye for an eye and a tooth for a tooth') and the associated idea, usual in penal law, that the punishment should be proportionate to the offence.

A second feature characterizing the efficiency and maximin norms -functions (1) and (2)- is that they are *non-consequentialistic*, that is, they prescribe actions without taking into account their *expected* material effects. To understand this point, consider a two-player sequential game where both the first and the second mover have available two choices (one consistent with the efficiency norm, the other not), and suppose that the first mover knows for sure that, if she complies, the second mover will then deviate, -in other words, she knows that a fairmax outcome will not be reached if she complies. Even in that case, the efficiency norm extols her to respect it (an analogous argument can be made with respect to the maximin norm). In this sense, these norms have arguably a striking similarity with Kant's Categorical Imperative -i.e., 'Act only on that maxim through which you can at the same time will that it should become a universal law'-, which is strongly non-consequentialistic.

The efficiency and maximin norms are non-consequentialistic because their prescriptions at any information set do not depend on how *other players are expected to play*.[6]

[6] Arguably, beliefs about future behavior are not the only important ones. In games of imperfect information, where players may not know exactly what other players (including Nature) did before, beliefs about past behavior might be also important for players to infer the consequences of their actions.

One way to relax non-consequentialism is assuming that the mover at h has probabilistic beliefs about future play so that any action $a \in A(h)$ induces a lottery $L(a)$ over the set of material outcomes $X(h)$. In this setting, and given any function F representing the players' expected social welfare (like, say, the expected social surplus), one could define an F-fairmax lottery at h as any lottery maximizing function F over the set of lotteries $\{L(a)\}_{a \in A(h)}$, and an F-fairmax action of h as any action $a \in A(h)$ inducing an F-fairmax lottery. In this setting, therefore, an action is F-fairmax at h if it induces a lottery of material outcomes that maximizes expected fairness, given the actor's beliefs at h. [7]

We could think of many other examples of norms. In particular, and although all previous examples of norms have something to do with distributive and retributive justice -how the 'social cake' is shared, how much effort each one should put in a task, the punishment that a transgressor should receive, etc.-, we stress that definition 1 can be used to formally represent qualitatively different norms, like (1) rules of etiquette -e.g., those indicating how the silverware should be set for dinner-, (2) norms regulating what one should eat -e.g., Muslims are told not to eat pork, Jews to eat *kosher*-, (3) norms prohibiting certain sexual practices in some societies -e.g., incest, sodomy, masturbation-, and (4) norms regulating communication -e.g., do not lie, do not commit blasphemy, do not use "dirty" language, etc. For examples of this kind of norms, the interested reader may consult López-Pérez (2007), where we provide formal examples of honesty norms.

3 Aversion to Norm-Breaking: Axioms

In the previous section we formally introduced norms and provided several examples. Here we move a step further and assume that norms can be internalized and hence shape preferences. More precisely, we introduce four axioms on the preferences of someone who has internalized a norm and can hence feel either guilt or shame when transgressing it.

Given any norm, let $R_{-i}(z)$ denote the set of players other than i that respected the norm in the history of z -i.e., chose always as commended-, and $I(z)$ denote an indicator that takes value 0 if player i respected the norm in the history of z and 1 if player i deviated. Observe that these two concepts are determined by the norm, and hence they are exogenous to the model (as the norm). Further, let $r_{-i}(z) \in [0,1]$ denote the proportion of players different than i that respected the norm in the history of z -i.e., the cardinality of $R_{-i}(z)$ divided by $n-1$. Note that a unique triple $[x_i(z), I(z), r_{-i}(z)]$ can be assigned to each z and for any player i.

[7] The prescriptions of a reciprocal norm, which depend on who violated the norm in the past, might be also conditional on beliefs about the set of previous norm respecters (see previous footnote).

If player i has internalized a norm, we postulate that her preferences over the set of terminal nodes of a game depend on $[x_i(z), I(z), r_{-i}(z)]$ -consequently, preferences are defined on the Cartesian product $\mathbb{R} \times \{0,1\} \times [0,1]$. Intuitively, the term $x_i(z)$ represents material joy while the two remaining terms correspond to the self-conscious emotions, whose activation depends on whether one complies with the norm -the term $I(z)$- and whose intensity is strongly correlated with others' compliant behavior -term $r_{-i}(z)$- for the cognitive reasons that we will mention later when studying the psychology of shame and guilt. Assuming that the preference ordering only depends on these three terms means, among other things, that the remorse from deviating does not depend on the harm done unto others or, more generally, on the material consequences of one's actions. This is admittedly unrealistic, but convenient at this stage of the analysis. Nevertheless, we suggest in the conclusion how to extend the model in this line.

To introduce the first axiom, consider two histories of play -i.e., two terminal nodes- along which player i always respected the norm. In this case, we posit that player i prefers the terminal node where his material payoff is higher. This is admittedly a gross simplification: Norm followers may be also concerned wether others respected the norm as well, since they may become angry or indignant if others deviate from the norm while they respect it. Nevertheless, if one isolates the effect of both material joy and self-conscious emotions on human motivation from that of other emotions, this axiom seems reasonable.

Axiom 1 (MI) *(Material interest) For any* $[a,0,r']$, $[b,0,r''] \in \mathbb{R} \times \{0,1\} \times [0,1]$,

$$[a,0,r'] \succsim_i [b,0,r''] \Leftrightarrow a \geqslant b.$$

One implication of axiom **MI** concerns *passive* players -i.e., players who make no choice in the game and that consequently can never be deviators. A passive player cannot bear responsibility for any transgression, and hence she cannot feel either guilty or ashamed of her actual play. Given that we disregard other emotions, it follows that her most preferred outcome is that in which she gets the largest material reward.

Our next axiom models the idea that transgressing an internalized norm triggers painful emotions. *Material payoffs being constant,* therefore, one strictly prefers to comply with an internalized norm than not to. This implies, for instance, that a judge who finds binding a norm of honesty and who believes defendant A to be very likely blameworthy, would rather condemn A if the material benefits of telling the truth are as large as those of lying -i.e., declaring A innocent.

Axiom 2 (IN) *(Internalization) For any* $a \in \mathbb{R}$, $r', r'' \in [0,1]$,

$$[a, 0, r'] \succ_i [a, 1, r''].$$

Note that axiom **IN** holds even if $r = 0$, that is, we assume that *ceteris paribus* a player who has internalized a norm strictly prefers to respect it even if no other player respects it. Some readers might find this contentious, and argue that the player should then be indifferent between respecting or transgressing the norm (this idea could be easily introduced in our model). However, we will provide one argument in favor of this hypothesis later, when discussing the psychology of guilt.

The third axiom indicates that people are able to trade off material interests and self-conscious emotions. A law of demand follows from this: The larger the expected price of obeying the norm, the less compliance there should be. Thus, the judge of our previous example may accept to declare A innocent if she is paid a high enough bribe or if she is credibly threatened to be killed otherwise.[8] As the previous axiom, this might appear to be a contentious axiom as well: As we explain in the next section, people are often unable to correctly forecast their emotional responses and hence to make the kind of trade off assumed here. Nonetheless, this does not pose a major problem if we understand that what people trade off with their expected material payoff is the emotion that they *expect* to feel if they deviate from the norm. This is an important implicit assumption of the analysis here.

Axiom 3 (TO) *(Trade off) For any* $a \in \mathbb{R}$ *and any* $r \in [0,1]$, *there exists* $a^* \in \mathbb{R}$ *such that*

$$[a, 1, r] \sim_i [a^*, 0, r].$$

In axiom **TO**, $[a^*, 0, r]$ is the *remorse-free equivalent* of $[a, 1, r]$. The difference $a - a^*$ measures the *minimum* net material payoff that the player should earn to be willing to violate the norm. The bribe to be paid to the judge to corrupt her is one example. It is easy to prove that such net payoff is always strictly positive.

Lemma 1 *If preferences* $\succsim_i$ *are rational and satisfy axioms* ***MI***, ***IN***, *and* ***TO***, *then* $a - a^* > 0$.

Proof. From axiom **IN**, it follows that $[a, 0, r] \succ_i [a, 1, r]$. Transitivity then implies $[a, 0, r] \succ_i [a^*, 0, r]$, whereas axiom **MI** gives finally $a > a^*$. ■

[8] Naturally, assuming that everyone's dignity has a price is compatible with heterogeneity. In this sense, a judge is unsubornable when the cost of buying her off is larger than the briber's material benefit of being declared innocent.

Our last axiom models a concern for personal status -more on this later. Intuitively, people compare with others when they transgress a norm -for any game, we implicitly assume that a player's reference group is composed by the other $n-1$ players- and they feel more badly as the proportion of *norm followers* increases. The judge of our example will feel especially badly if she believes to be the only corrupt judge in the court whereas her pain will be alleviated if she believes that most judges accept bribes.

Axiom 4 (S) *(Status) If* $[a,1,r] \sim_i [a^*,0,r]$ *and* $[b,1,r'] \sim_i [b^*,0,r']$ *then*

$$r' \geqslant r \Leftrightarrow b - b^* \geqslant a - a^*$$

Consider now any rational preference profile satisfying axioms 1 to 4. We now prove that such ordering can be represented by the utility function that we assumed in López-Pérez (2008).

Theorem 1 *A rational preference ordering* $\succsim_i$ *satisfying axioms* ***MI***, ***IN***, ***TO***, *and* ***S*** *admits the following functional representation*

$$U_i[x_i(z), I(z), r_{-i}(z)] = x_i(z) - I(z) \cdot \gamma(r_{-i})$$

where $\gamma : [0,1] \to \mathbb{R}^+$ *is a strictly increasing function.*

Proof. *Consider any triple* $[a,1,r]$ *and its remorse-free equivalent* $[a^*,0,r]$. *Lemma 1 indicates that* $a - a^* > 0$, *whereas axiom* ***S*** *implies that* $a - a^*$ *only depends on* r *and, moreover, it is strictly increasing in* r. *Given this, function* $\gamma(r)$ *is taken to be equal to* $a - a^*$.

$\Leftarrow$ *Take two alternatives* $[a,1,r],[b,1,r']$ *and assume first that* $U_i[a,1,r] \geqslant U_i[b,1,r']$. *Hence,* $a-(a-a^*) \geqslant b-(b-b^*) \Rightarrow a^* \geqslant b^*$ *so that it follows from axiom* ***MI*** *that* $[a^*,0,r] \succsim_i [b^*,0,r']$, *whereas transitivity implies then* $[a,1,r] \succsim_i [b,1,r']$. *The same line of reasoning proves for any other pair of alternatives that* $U_i[x_i(z),I(z),r(z)] \geqslant U_i[x_i(z'),I(z'),r(z')]$ *implies* $[x_i(z),I(z),r(z)] \succsim_i [x_i(z'),I(z'),r(z')]$.

$\Rightarrow$ *Take two alternatives* $[a,1,r],[b,1,r']$ *and consider first the case* $[a,1,r] \succsim_i [b,1,r']$, *which implies* $[a^*,0,r] \succsim_i [b^*,0,r']$. *In turn, axiom* ***MI*** *implies* $a^* \geqslant b^* \Rightarrow U_i[a,1,r] \geqslant U_i[b,1,r']$. *A similar argument shows that* $[x_i(z),I(z),r(z)] \succsim_i [x_i(z'),I(z'),r(z')]$ *implies* $U_i[x_i(z),I(z),r(z)] \geqslant U_i[x_i(z'),I(z'),r(z')]$ *in the remaining cases.* ■

We make a number of remarks on the previous result. *First*, note that it does not restrict much the properties of function $\gamma(r)$, apart of being strictly positive and increasing on r. Indeed, axioms 1 to 4 are too weak to get a more precise utility function, and

additional hypotheses should be introduced for this. One sensible assumption might be strict convexity: $\gamma(r' + \frac{1}{n-1}) - \gamma(r') > \gamma(r + \frac{1}{n-1}) - \gamma(r)$ if $1 > r' > r$. This assumption states that the intensity of a deviator's pain grows at an *increasing* rate as the proportion of players who respect the norm rises and that, consequently, people care relatively few about being unprincipled if most of the others are also unprincipled. *Second*, observe that the utility characterization is not unique. Clearly, any monotone transformation of our utility function can also represent a preference ordering satisfying axioms 1 to 4, and other utility functions might play this role as well. We do not pursue this point further, however, as our main objective was showing that the utility model in López-Pérez (2008) is in line with our axiomatic analysis. Indeed, the utility function posited in that model has the structure $x_i(z) - I(z) \cdot \gamma(r_{-i})$.[9]

Finally, the previous theorem assumes that preferences are rational -i.e., complete and transitive-, and the reader may wonder whether a preference ordering satisfying our four axioms can at the same time be rational. To start, it is rather clear that completeness does not contradict any of our axioms -these axioms assume that the ordering of alternatives has certain properties, not that the ordering does not exist. With respect to transitivity, it suffices to mention one preference ordering that satisfies transitivity and our four axioms. For this, consider a function that assigns number $b - G - r$ $(G > 0)$ to any triple $[b, 1, r]$, and number c to any triple $[c, 0, r']$, and consider a complete ordering such that any triple is strictly preferred than another one with a lower number (and indifferent to any triple with the same number). Obviously, this ordering satisfies axiom **MI**. It also satisfies axioms **IN** (this is true even if $r = 0$, as condition $G > 0$ guarantees), **TO** (implicitly, difference $a - a^*$ equals $r + G$ for any triple $[a, 1, r]$ and its remorse-free equivalent $[a^*, 0, r]$ in this ordering), and **S**. To finish, transitivity also holds because this preference ranking is based on the ranking of the real numbers, which is indeed transitive. To sum up, rationality is compatible with our four axioms.

4 The Psychology of Guilt and Shame

In the prior section we posited that the preferences of an agent who has internalized a norm must satisfy four axioms. In this section we review some of the psychological literature on

[9] To be precise, the model in López-Pérez (2008) assumes for simplicity that $\gamma(0) = 0$. This assumption is at odds with axiom **IN**, but could be easily relaxed and does not affect our results in López-Pérez (2008), provided that $\gamma(0)$ is small. Additionally, López-Pérez (2008) define $r_{-i}(z)$ as the cardinality of set $R_{-i}(z)$, and not as its overall proportion, as we do here. Both assumptions lead to qualitatively similar predictions in the games analyzed in that paper (since we focus there on two-player games) and do not change much our analysis here. We have introduced this new specification in our model here because we now consider that it is more empirically relevant in multiple-player games.

shame and guilt, and argue that the intuition behind our axioms (in particular axioms **IN** and **S**) lies in the psychology of these two emotions. Additionally, we aim to clarify the difference between guilt and shame, as we feel that this point is sometimes misunderstood -for instance, one can frequently read that guilt is 'internal' or 'private' while shame is 'external' or 'public'; however this view is not supported by the most recent psychological literature.

4.1 Shame: Aspiration Levels and the Perceived Self

We start by introducing two key psychological concepts. For this, we assume that personal traits -either bodily ones like stature, color of the eyes/hair, weight, etc., or spiritual ones like courage, creativity, intelligence, moral integrity, wit, etc.- can be somehow measured or evaluated. In this case, the *level of (self-) aspiration* for a certain trait Q is the level of Q that one considers 'good enough', while one's *perceived self* (or simply, one's *self*) with respect to Q is the level of Q that one believes to have attained, that is, the image that one has of herself on that regard.

Needless to say, the aspiration level and the perceived self need not coincide for every trait. Thus, someone may aspire to be slim although she is actually obese. Another person may wish to be witty but believes to be rather dull. We posit that shame of a certain trait is activated when there exists such a negative gap between the aspiration level and the perceived self on that trait (conversely, a positive gap should trigger pride). Note well that the term 'shame' refers here to any emotional feeling triggered by that negative gap. This includes proper shame, but also embarrassment and low self-esteem.

This 'negative gap' hypothesis is consistent with the psychological literature, which often considers shame to be triggered by cognitions regarding the self (Lewis, 1971; Lewis, 1992; Lindsay-Hartz et al., 1995; Tangney and Dearing, 2002). Thus, Lewis (1971, 30; emphasis in the original) claims that "The experience of shame is directly about the *self*, which is the focus of evaluation." Further, evidence from questionnaires indicates that subjects often report the desire to undo some aspect of the self when they are asked to narrate a shame episode, thus setting into motion thoughts like "If only I weren't such-and-such kind of person" -see Tangney (1995, 117) and references therein.

Additional evidence in support of the 'negative gap' hypothesis can be obtained by studying how people evaluate the perceived self and the level of aspiration. Festinger (1954) provide evidence and a theory on how the perceived self is created, noting first that people evaluate their own personal qualities through objective means if available -for instance, bodily qualities like one's height or weight admit such nonfaulty evalua-

tion. Otherwise, a person resorts to others' opinions about her personal qualities.[10] For instance, the belief that one is beautiful heavily depends on whether others believe such thing -of course, one might weight differently each person's opinions.

A possible reading of Festinger (1954) also suggests a theory on how people construct their aspiration level. Provided that an agent can evaluate the level of trait Q of any member of her reference group (either directly or by inference from his behavior), Festinger suggests that the agent's level of aspiration for Q depends on that information:

> "Level of aspiration experiments have been performed where, after a series of trials in which the person is unable to compare his performance with others, there occurs a series of trials in which the person has available to him the knowledge of how others *like himself* performed on each trial [...]. When the 'others like himself' have scores different from his own, his stated 'level of aspiration' (his statement of what he considers is good performance) almost always moves close to the level of performance of others. [...] When the reported performance of others is about equal to his own score, [...] his level of aspiration shows little variability."[11]

In addition, internalized standards and norms also seem to play an important role in determining the level of aspiration. For instance, the level of tax compliance that a taxpayer wishes to accomplish might depend on the assessed level of overall tax evasion -a signal of the others' honesty- but also on whether she considers taxpaying a civic duty. As another example, a teenager's desired bodyweight might be a function of the weight of her relatives and friends, but also of the beauty standards that the media spread.

[10] This raises the question of where others' opinions come from. Festinger does not provide a clear answer to this.

[11] Festinger (1954, 140); observe the implicit mention to the reference group (emphasis in the original).

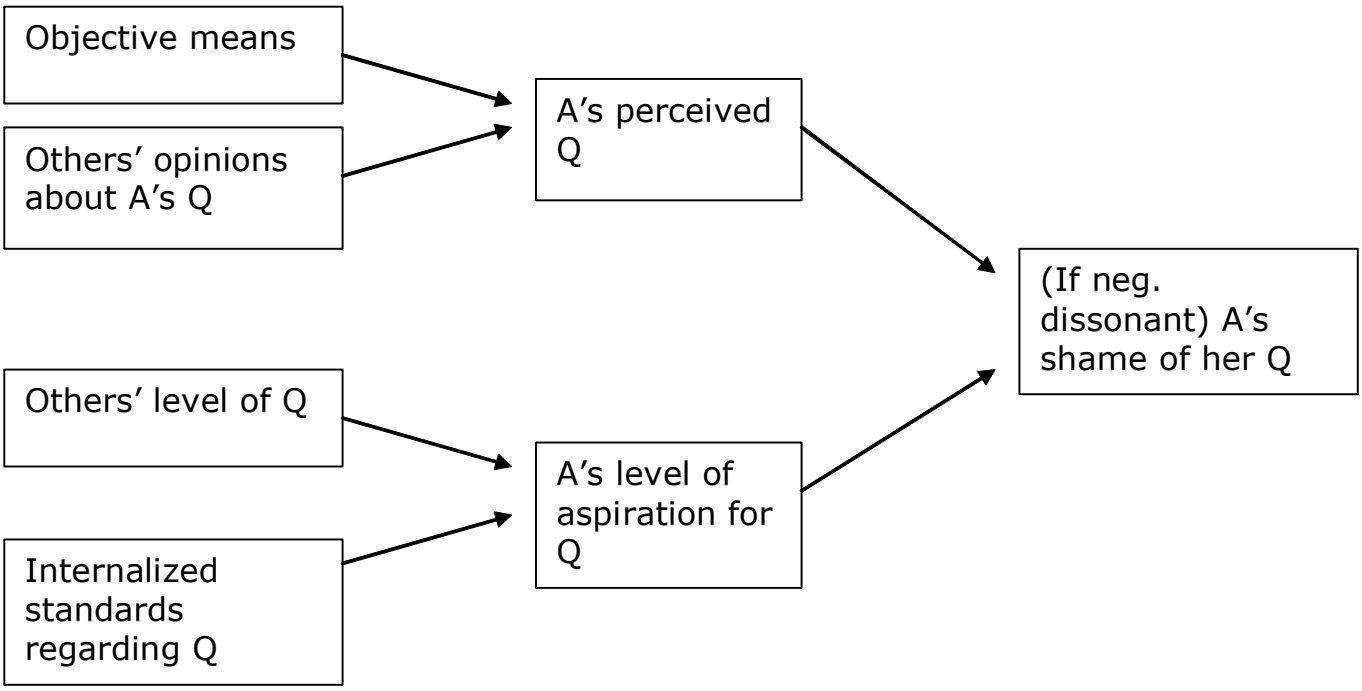

Figure 2: The Genesis of Shame

Figure 2 summarizes the previous ideas and shows how shame is activated (arrows indicate causality).[12] To start, and since internalized norms influence our level of aspiration, a deviation from those norms is likely to make us feel ashamed (axiom **IN**). In addition, the aspiration level also depends on how one's behavior (and the qualities that this behavior signals) compares with others'. For this reason, a transgression is more likely to trigger (or intensify) shame if a significant *proportion* of the others do respect the norm.[13]

In other words, a sense of inferiority is likely to translate into stronger shame (axiom **S**) and hence shame intensity is highly conditional on what (we believe that) others have or do.[14] This can be illustrated with several examples.[15] Consider a person who regards littering or smoking at public places as improper behavior, and compare her feelings when she litters a park grassfield that is (i) clean or (ii) dirty; or if she smokes in a public place

[12]The figure suggests that the activation of shame is positively correlated with variables like social disapproval -i.e., bad opinions from the others- and public exposure (which can provide vivid information about others' disapproval of our behavior). We leave the study of these variables for further research.

[13]Note that we used the word proportion: If a participant in an IQ test ranks 100th out of a group of 100 people, that is much worse that if he ranks 100th out of a group of 1000 people. Nevertheless, it also seems that absolute numbers matter: A low position within a very small reference group might be less hurtful than a low position within a big group, even if the relative position -measured, say, by the decile one occupies- is the same in both cases. For simplicity, our model focuses on the 'proportion' effect.

[14]Therefore, shame might be important to enlighten why relative status is key for happiness (Veblen 1934; Frank, 1985).

[15]By stressing how shame critically depends on others' behavior, these examples hint the role that Game Theory should play in its study.

where (i) nobody smokes or (ii) many people smoke. It is also illustrative that many citizens of countries where corruption and tax-evasion are widespread tend to justify their behavior by claiming 'If nobody complies, why should I comply?' This suggests (in line with axiom **S**) that the remorse from deviating dwindles as less people respect the norm: "Bad company ruins good morals." (Bible, 1 Corinthians 15:33).

All this is consistent with the psychological literature. Thus, Lewis (1971) stress that shame is typically associated to inferiority feelings, a sense of shrinking or of "being small", and Tangney and Dearing (2002) report much questionnaire evidence showing a strong correlation between a sense of inferiority and shame -consult also Lewis (1992), or Lindsay-Hartz et al. (1995). In addition, some sociobiological evidence *seems* to point in this direction as well. In effect, studies with primates show that dominant vervet monkeys have larger concentrations of the neurotransmitter serotonin in their blood than nondominant monkeys -i.e., those most likely to feel inferior- and that these differences are the result, rather than the cause, of the monkeys' dominance relationships -see Frank (1985, 23-28) for a review. The fact that inferiority feelings induce low serotonin levels is interesting because some evidence correlates low serotonin levels with mania and depression (the popular antidepressant Prozac probably acts by increasing the availability of serotonin), two phenomena closely linked to shame and low self-esteem -Lewis (1971), Lewis (1992) and Tangney and Dearing (2002).

4.2 Discriminating between Shame and Guilt

Consistent with our axiom **IN**, many researchers from different fields (Lewis 1971, Lewis 1992, Elster 1999, Posner and Rasmusen 1999, Tangney and Dearing 2002, Bowles and Gintis 2003) contend that guilt is activated when one transgresses an internalized norm, a characterization that is also similar to Freud's (in *Civilization and its Discontents*, guilt constitutes the punishment that the superego imposes on the ego when the latter follows id impulses and transgresses internalized standards).

However, we note that this *transgression hypothesis* is somewhat contentious. To start, some psychologists argue that the mere existence of advantageous inequities may elicit guilt, even if the outcome is not the result of some personal wrongdoing or, indeed, of any personal behavior at all. The phenomenon of *survivor guilt*, as it is called, has been documented in survivors of Hiroshima and the Holocaust, homosexual men who have tested *negative* for the human immunodeficiency virus, and people who keep their jobs when co-workers are fired, among others -see Baumeister (1994) for references. Many times, however, it happens that although neutral observers think the opposite, survivors believe that they could have done something to avert a fatal outcome -Elster (1999, 151).

Hence, this apparent survivor guilt can be often rationalized and understood in the typical framework.

Leaving aside the issue of survivor guilt, however, the transgression hypothesis faces a more puzzling problem. As internalized norms shape (at least partially) the aspiration level -recall Figure 2-, transgressing a norm should create a negative gap between the aspiration level and the self, and hence guilt should be indistinguishable from shame. Nevertheless, the influential account by Lewis (1971, 30) notes that the main difference between shame and guilt is that "In guilt, the self is not the central object of negative evaluation but rather the *thing* done or undone is the focus" (italics in the original). Why is the self not heavily impaired when one feels guilty? One might think of at least three possible reasons for this.

The first reason might be that, contrary to shame, guilt is not linked to feelings of inferiority so that the negative gap is minor. This could incidentally explain why, as Lewis (1992) and Tangney and Dearing (2002) note, guilt is usually *less* intense than shame, and provides a reason why axiom **IN** of our preference model holds even in the case $r = 0$. What explains this absence of inferiority feelings when feeling guilty? On one hand, the reference group of the transgressor might be empty, that is, she might not compare herself with others. Needless to say, this cannot occur as a result of an intentional effort: Attempting not to think about someone is a self-defeating task. What happens is simply that no image of the others comes to one's mind when deviating. Solitary transgressions might constitute propitious settings for this because transgressors lack then vivid information about the others -Baumeister et al. (1994, 251) report that people may feel guilty for failing at dieting, neglecting their studies, failing to exercise, having drugs, masturbating, and looking at nude pictures.[16] On the other hand, a transgressor might not feel inferior if everybody in her reference group deviates as well. As an aside, this point suggests that the *same* deviating behavior may elicit either guilt or shame depending on how others behave: A soldier that deserts from his unit in the middle of the battle will feel ashamed if he believes that fighting is his duty *and* most of his comrades fight, but she is likely to feel just guilty when everybody abandons the lines.[17]

The two other possible reasons why the self is not impaired in guilt could be that people believe that 'they are not what they do' when feeling guilty -that is, they think that their actions do not signal anything about their respective selves- or that they believe that the

[16] Taylor (1985, 88) notes that "I may feel guilty because I watch that silly television serial rather than improve my mind by reading great literature". In contrast, we might feel ashamed and not guilty if the thought that *others* despise that serial and never watch it came to our mind.

[17] This might explain why, as Tangney and Dearing (2002, 17) report, "Our analysis of the personal shame and guilt experiences provided by both children and adults indicate that there are very few, if any, 'classic' shame-inducing or guilt-inducing situations [i.e., behaviors]."

self changes, so that the self that chose to transgress the norm is different than the self that feels guilty later. June Tangney and her colleagues have tested these two theories -see Tangney and Dearing (2002, 66-69) for a summary- by means of two scales respectively assessing the belief that the self and behavior are congruent versus incongruent and the belief that the self is fixed versus malleable. 175 undergraduates were asked to complete those scales as well as a number of other measures including one scale of proneness to guilt. The authors report (i) no significant correlation between proneness to guilt and self-behavior congruence beliefs, and (ii) a negative correlation between proneness to guilt and the belief that the self is fixed. Point (i) suggests that guilt has nothing to do with believing that 'people are not what they do', whereas point (ii) hints that guilt-prone people might tend to perceive the negative gap as *transient or unstable*.

To illustrate this latter point, consider a young man who cheats his girlfriend for the first time in his life and feels very badly for that. He may find the intensity of that pang surprising and unexpected. In fact, he may even think that he would not have cheated her girlfriend had he understood how badly he was going to feel, and decide as a result not to cheat her never again.[18] In this sense, the damage to the self-image is transitory: Although he behaved badly, he believes not to be a bad person anymore -this belief, of course, may have fragile foundations although this is unsubstantial for the argument here. More generally, people might believe that the damage is transitory because they explain their norm transgression as a result of a preference that has changed afterwards, maybe because of a learning process.

That emotions are a source of preference reversals is consistent with much experimental evidence indicating that people predict rather inaccurately their future or imaginary emotional states (recall our discussion of axiom **TO**). Thus, Loewenstein and Lerner (2003, 629) report that

> "When people are in a 'cold' state -for example, not hungry, angry, sexually aroused, and so forth- they underappreciate what it will feel like to be in a hot state in the future and how such a state will affect their behavior. They make an analogous mistake when in a hot state and predicting how they will feel or behave when the heat dissipates -that is, when they are in a cold state. Such 'hot/cold empathy gaps' occur not only prospectively, when people predict their own future feelings and behavior, but also retrospectively [...]".

Note also that norm transgressions are sometimes the result of mistakes. Suppose, for instance, that someone makes an inappropriate comment about other person's physical

[18] Incidentally, guilt is associated to some action tendencies (Frijda, 1986), like confession and reparative behavior, that one is unable to forecast before commiting the bad deed.

appearance without any bad intention. She might feel badly for that, but that bad feeling may be attenuated by the belief that she has learnt not to do that again -if she has not committed repeatedly that mistake in the past, she may be confident that next time she will be more careful regarding what she says. Again, the transgression is only a signal of a transient dissonance and hence the self is not impaired.

To sum up, guilt seems to be activated by one's improper behavior when, in contrast to shame, that behavior is interpreted as signalling a *transient* and relatively less intense negative gap. In turn, transience and low emotional intensity might be respectively explained by preference reversals (or because the improper behavior was the result of a mistake) and the absence of inferiority feelings.

We finish this section by comparing the previous distinction between guilt and shame with a classical anthropological and psychological view that understood guilt as a private affair and shame as somewhat the correlate when our deviation is subject to public exposure. Buss (1980, 159), for instance, state that "The best test of guilt is whether anyone else knows of the transgression. In true guilt, no one need know. [...] Shame is essentially public; if no one else knows, there is no basis for shame." Many psychologists find this view rather inaccurate. Although there is indeed some correlation between public exposure and shame activation (see footnote 12), evidence from questionnaires suggests that shame is not restricted to public settings. Thus, Tangney and Dearing (2002, ch. 2) report results of narrative accounts of personal shame and guilt experiences and stress that a substantial number of these events occurred when the person was *alone* and that, importantly, solitary shame experiences were just as common as solitary guilt experiences. Introspection is also at odds with the public-private distinction. For instance, one may be ashamed by his own *private* thoughts and desires as may happen, for instance, when thinking about one's friend's husband or wife in a sexual way or when desiring someone's death. In addition, people may feel ashamed when lying or behaving unfairly, even if they do not expect being caught. In summary, one does not need to be exposed to the others to have low self-esteem.

5 Conclusion

Social researchers have constantly remarked that the values and norms that an individual shelters are crucial to understand her conduct and, more generally, the performance of those societal groups that share them. Nevertheless, it is not enough to point out that someone has certain norms to explain why she follows what they dictate; it is also fundamental to understand the motivational forces that drive her. We believe that self-

conscious emotions like guilt and shame are important in this regard, and have offered a formal analysis here of how they affect preferences.

A series of questions deserve further study. *First*, our analysis here assumes for parsimony that the pain of breaking a norm does not depend on the material consequences of the transgressor's behavior. This is indeed a simplification of the true thing: If one person drives recklessly and an accident occurs, he will feel much more badly if someone dies as a result.[19] If the norm is a norm of distributive justice with associated function F (see section 2), the pain might depend on the difference between the social welfare $F(z_o)$ at the optimal node z_o, and that at the actual node z_A.

Second, we could include additional variables that affect shame and guilt -for instance, Baumeister et al. (1994) emphasize *social distance*, that is, bad feelings may be relatively more intense in close relationships involving relatives or friends; and Battigalli and Dufwenberg (2007) assume that people feel the more guilty the more they disappoint another people. *Third*, we argued that an unfavorable comparison with the others often triggers shame. However, who are the others? That is, who do we compare with? We hypothesized that people compare with those who directly interact with them -i.e., with the set of players. This might be rather accurate in certain laboratory settings, but in general it seems a gross simplification. As an extreme example, dead relatives might be part of one's reference group. *Fourth*, research on the cognitive processes that involve the formation of the level of aspiration and the perceived self is crucial too. Finally, a key question is why people internalize norms. The answer, as Bowles and Gintis (2003, 20) stress, might have to do with *bounded rationality* and evolution:

> "The internalization of norms eliminates many of the cost/benefit calculations and replaces them with simple moral and prudential guidelines for action. Individuals who internalize norms may therefore have higher payoffs than those who do not, so the psychological mechanisms of internalization are evolutionarily selected."

References

[1] Akerlof, George A., and Rachel E. Kranton, 2005. "Identity and the Economics of Organizations." Journal of Economic Perspectives, 19(1), 9-32.

[2] Arrow, Kenneth J., 1974. *The Limits of Organization*. Norton & Company.

[19] Note that one of the action tendencies of guilt is to repair -Lewis 1971, Tangney and Dearing 2002. This behavior might be a way to reduce the harm done unto others and thus lessen the burden of guilt.

[3] Battigalli, Pierpaolo, and Martin Dufwenberg, 2007. "Guilt in Games." American Economic Review, Papers & Proceedings, 97, 170-176.

[4] Baumeister, Roy F., Arlene M. Stillwell and Todd F. Heatherton, 1994. "Guilt: An Interpersonal Approach." Psychological Bulletin, 115(2), 243-267.

[5] Becker, Gary (1996). *Accounting for Tastes*. Harvard University Press.

[6] Beer, J. S., E. A. Heerey, D. Keltner, D. Scabini, and R. T. Knight, 2003. "The Regulatory Function of Self- Conscious Emotions: Insights from Patients with Orbitofrontal Damage." Journal of Personality and Social Psychology, 85(4), 594-604.

[7] Bosman, Ronald and Frans van Winden, 2002. "Emotional Hazard in a Power-to-Take Experiment." Economic Journal, 112, 147-69.

[8] Bowles, Samuel and Herbert Gintis, 2003. "Prosocial Emotions." Mimeo.

[9] Buss, A., 1980. *Self-Consciousness and Social Anxiety*. San Francisco: Freeman.

[10] Conlin, Michael, Michael Lynn, and Ted O'Donoghue, 2003. "The Norm of Restaurant Tipping." Journal of Economic Behavior and Organization, 52, 297-321.

[11] Damasio, Antonio, 1994. *Descartes' Error*. New York: Putnam.

[12] Elster, Jon, 1989. *The Cement of Society*: *A Study of Social Order.* Cambridge University Press.

[13] Elster, Jon, 1999. *Alchemies of the Mind. Rationality and the Emotions.* Cambridge University Press.

[14] Fehr, Ernst and Urs Fischbacher, 2004. "Social Norms and Human Cooperation." Trends in Cognitive Sciences, 8(4), 185-190.

[15] Fehr, Ernst and Simon Gächter, 2000. "Fairness and Retaliation: The Economics of Reciprocity." Journal of Economic Perspectives,14(3), 159-181.

[16] Festinger, Leon, 1954 [1989]. "A Theory of Social Comparison Processes". Human Relations 7, 117-40. Reprinted in Stanley Schachter and Michael Gazzaniga, eds., *Extending Psychological Frontiers. Selected works of Leon Festinger*. New York: Russel Sage Foundation.

[17] Frank, Robert H., 1985. *Choosing the Right Pond. Human Behavior and the Quest for Status.* New York, Oxford University Press.

[18] Frijda, N., 1986. *The Emotions.* Cambridge University Press.

[19] Homans, George C., 1961. *Social Behavior: Its Elementary Forms.* New York: Harcourt, Brace.

[20] Horne, Christine, 2001. "Sociological Perspectives on the Emergence of Social Norms." In Michael Hechter and Karl Dieter Opp (eds.), *Social Norms.* New York: Russell Sage Foundation, 3-34.

[21] Kandori, Michihiro, 1992. "Social Norms and Community Enforcement," Review of Economic Studies, 59, 63-80.

[22] Lewis, H. B., 1971. *Shame and Guilt in Neurosis.* New York: International Universities Press.

[23] Lewis, Michael, 1992. *Shame: The Exposed Self.* New York: The Free Press.

[24] Lindsay-Hartz, J., J. de Rivera, and M. F. Mascolo, 1995. "Differentiating Guilt and Shame and Their Effects on Motivation." in June P. Tangney and Kurt W. Fischer (eds.), *Self Conscious Emotions.* New York: Guilford Press.

[25] Loewenstein, George and Jennifer S. Lerner, 2003. "The Role of Affect in Decision Making." in R. J. Davidson, K. R. Scherer and H. H. Goldsmith, eds., *Handbook of Affective Sciences*, Oxford, UK: Oxford University Press.

[26] López-Pérez, Raúl, 2008. "Aversion to Norm-Breaking: A Model." Games and Economic Behavior, 64, 237-267.

[27] López-Pérez, Raúl, 2007. "The Power of Words: Why Communication Fosters Cooperation and Efficiency". Mimeo.

[28] Mealy, Linda, 1995. "The Sociobiology of Sociopathy: An Integrated Evolutionary Model." Behavioral and Brain Sciences, 18(3), 523–599.

[29] Parsons, Talcott, 1937. *The Structure of Social Action.* New York: McGraw-Hill.

[30] Posner, Richard A., and Eric B. Rasmusen, 1999. "Creating and Enforcing Norms, with Special Reference to Sanctions." International review of Law and Economics, 19, 369-82.

[31] Tadelis, Steve, 2008. "The Power of Shame and the Rationality of Trust." Mimeo.

[32] Tangney, June P., 1995. "Shame and Guilt in Interpersonal Relationships." In June P. Tangney and Kurt W. Fischer (eds.), *Self Conscious Emotions*. New York: Guilford Press, 114-139.

[33] Tangney, June P. and Ronda L. Dearing, 2002. *Shame and Guilt.* New York: Guilford Press.

[34] Taylor, C., 1985. *Pride, Shame and Guilt.* Oxford University Press.

[35] Veblen, Thorstein, 1934. *The Theory of the Leisure Class*. New York: Modern Library.

[36] Voss, Thomas, 2001. "Game-Theoretical Perspectives on the Emergence of Social Norms." In Michael Hechter and Karl Dieter Opp (eds.), *Social Norms*. New York: Russell Sage Foundation, 105-136.